International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-2, March 2014

Bilingual Information Retrieval System For English And Marathi

Pratik Pokharkar,Onkar Nikam, Anurag Mishra, Divya Raisinghani

***Abstract*— Our system addresses the design and implementation of BiLingual Information Retrieval system on the domain, Festival.It is built for Marathi language working with the same efficiency.**
According to User's query ,searching,translation and information extraction is done effectively. The main task is to retrieve the solution for the user typed query in the both the languages one that of the query and the standard English language. In this process, a Ontological tree is built for the domain in such a way that there are entries of important keywords in the above listed two languages in every node of the tree. A Part-Of-Speech (POS) Tagger is used to divide the sentence into words and assign their POS then determine the keywords from the given query. Based on the context, the keywords are translated to appropriate languages using the Ontological tree. A search is performed and documents are retrieved based on the keywords. With the use of the Ontological tree, Information Extraction is done. And at last, the solution for the query is translated back to the query language and produced to the user as per his requirement.

***Keywords*-Information-Retrieval,Information extraction, Ontology tree,POS Tagger.**

I. INTRODUCTION

In query translation approach the query terms can be a single word or a phrase. So, we need a combination of phrase reorganization, pattern-based phrase translation and query expansion before and after translation to improve dictionary-based query translation. The task is to retrieve the solution for the user given query in the same language as that of the query. In this process, a Ontological tree is built for the domain in such a way that there are entries of important keywords in the above listed two languages in every node of the tree. A Part-Of-Speech () Tagger is used to divide the sentence into words and assign their POS then determine the keywords from the given query. Based on the context, the keywords are translated to appropriate languages using the Ontological tree. A search is performed and documents are retrieved based on the keywords. With the use of the Ontological tree, Information Extraction is done. Finally, the solution for the query is translated back to the query language and produced to the user.

Manuscript received March ,23 2014.

Pratik M Pokharkar, Computer Engineering, Pimpri Chinchwad College Of Engineering, Pune, India, 9975434540, (e-mail: pratik.pokharkar2611@gmail.com).
Onkar Nikam, Computer Engineering, Pimpri Chinchwad College Of Engineering, Pune, India, 7709995956, (e-mail: nikamonkar56@gmail.com.
Anurag Mishra, Computer Engineering, Pimpri Chinchwad College Of Engineering, Pune, India, 9975121198, (e-mail: anurag.s.mishra014@gmail.com).
Divya Raisinghani, Computer Engineering, Pimpri Chinchwad College Of Engineering, Pune, India, 9029449238, (e-mail: divyaraisinghani0507@gmail.com).

II. LITERATURE REVIEW

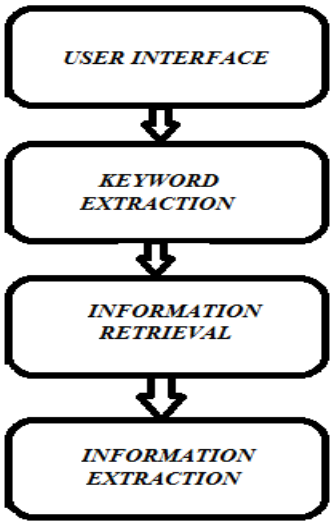
- **Cross Language Information Retrieval :**

Query translation has emerged as the most popular technique for CLIR typically achieving between and of the retrieval effectiveness that is reported for comparable monolingual techniques when coupled with simple linguistic processing such as part of speech tagging or phrase indexing Query translation strategies are relatively efficient when short queries are presented but a lack of adequate linguistic context in queries containing only a few words may limit the ability of systems to select the most appropriate translations for the query terms Machine translation systems seek to exploit contextual clues in full length documents to produce the best possible translations and it is an open question whether a retrieval system based on automatic machine translation of each document can outperform query translation. With the explosion of online non-English documents, cross language information retrieval (CLIR) systems have become increasingly important in recent years. Research in the area of CLIR has focused mainly on methods for query translation. In particular, dictionarybased translation has been a commonly used method because of its simplicity and the increasing availability of machine readable bilingual dictionaries. However, besides the problem of completeness of the dictionary, we are also faced with the problem of ambiguity

in translation, i.e. the selection of the correct translation word(s) from the dictionary.

Modules of proposed System:

- User Interface :- User enters the query in native language.
- Keyword Extraction:- in this module the query is processed by POS tagger ,where query is divided into words, and important keywords are extracted.
- Information Retrieval:- Using Ontological approach query is translated.
- Information Extraction:- In this module the relevant pages are Extracted and displayed to the user.



III . PROPOSED METHODOLOGY

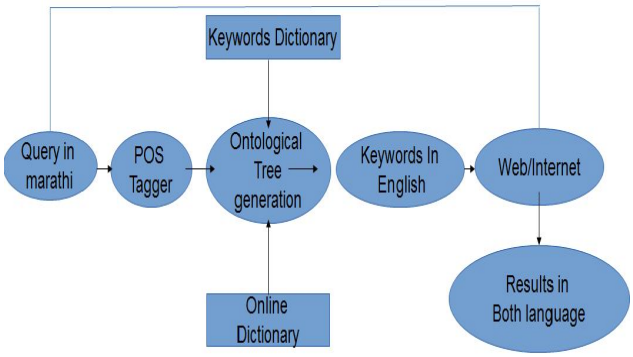


FIGURE 1. ARCHITECTURE DIAGRAM

Ontology is being used to extract the concepts from the documents and queries. According to the query terms, it is translated to either language. Phrase translation is being used instead of word translation once it is recognized. Query extension may be before or after translation. User enters the query in Marathi language through virtual keyboard. This query is passed to POS Tagger, It divide the query in to words and assigns parts of speech to each words. After tagging it is passed to ontological tree and appropriate translation of query into English language is done, using word mapping in ontological database.After translation page rank algorithm is used to dispaly relevant pages of the query in both languages.

The steps for Bilingual searching are:

Extract all synsets from WordNet for entered word. If for any word more than one synset exists then calculate semantic similarity for this word along with nearby words. Select most related synsets for the entered word based on semantic similarity. All the synonyms, all the hyponyms and one hypernym are added for the entered word.

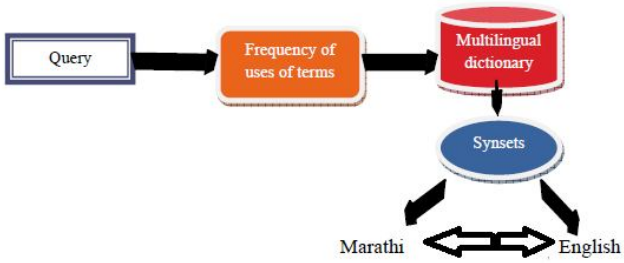


Figure : Use of Synsets

POS Tagger:-

Pos tagger is Used to Divide the Sentence into words.It assigns Parts Of Speech to each Word.Also Called as Grammatical Tagging.
For eg. I am an Engineer. Will convert into ->
I/PRP am/VBP an/DT Engineer/NNP J

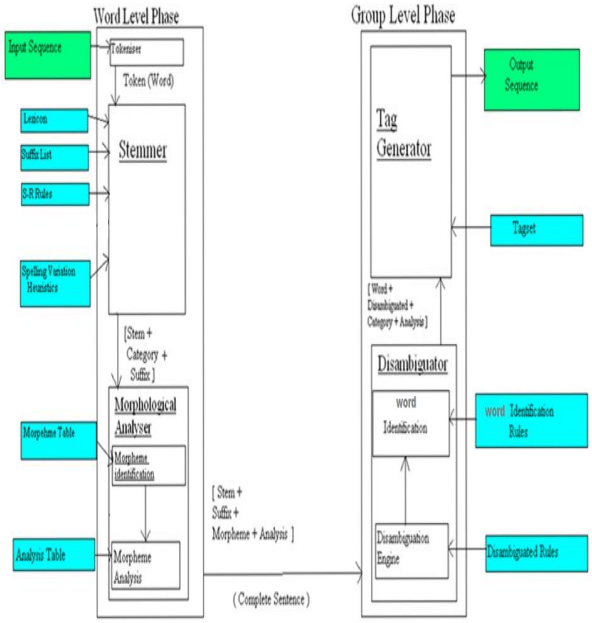


Figure... Design Model For POS Tagger

Ontological Tree:-

Ontological Tree is defined as grouping of entities.It is grouping of entities based on similarities and differences. Ontological tree consists :
1) Nodes with multiple entries.
2) Entry for both the languages in every node.

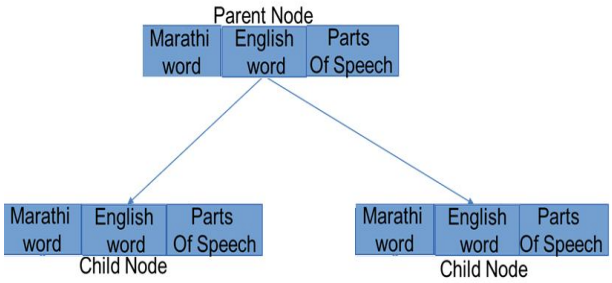


Figure: Structure For Ontological Tree

IV. FUTURE SCOPE

Suppose In future ,If we decide to Increase the scope of the project i.e we have to add one or more languages then we have to just add one field in the ontology tree for that language.

V. CONCLUSION

Hence we have studied and design the system in which query which is enter by user in Marathi language is translated and relevant pages are displayed in both Marathi an English languages. Also we have studied concepts of POS Tagger, Ontological approach for translation and storing data.

REFERENCES

[1] BiLingual Information Retrieval System for English and Tamil - Dr.S.Saraswathi, Asma Siddhiqaa.M, Kalaimagal.K, Kalaiyarasi.M
[2] Cross-Language Information Retrieval: the way ahead - Fredric C. Gey , Noriko Kando, Carol Peters
[3] A Personalized Ontology Model for Web Information Gathering - Xiaohui Tao, Yuefeng
Li, and Ning Zhong, Senior Member, IEEE
[4] Indic Language Translation in CLIR Using Virtual Keyboard - Mallamma V. Reddy and M. Hanumanthappa


Pratik M Pokharkar
Pimpri Chinchwad College Of Engineering
B.E. Computer Engineering


Divya Raisinghani
Pimpri Chinchwad College Of Engineering
B.E. Computer Engineering


Anurag Mishra
Pimpri Chinchwad College Of Engineering
B.E. Computer Engineering


Onkar C. Nikam
Pimpri Chinchwad College Of Engineering
B.E. Computer Engineering